Ability of the misfolding avoidance hypothesis to explain protein evolutionary rates in the light of empirical evidence

Germán Plata[1,2], Dinara R. Usmanova[1] and Dennis Vitkup[1,3†]

[1]Department of Systems Biology, Columbia University, 1130 Saint Nicholas Ave., New York, NY 10032, USA. [2]Present address: Elanco Animal Health, 2500 Innovation Way, Greenfield, IN 46140, USA. [3]Department of Biomedical Informatics, Columbia University, 622 West 168th street, New York, NY 10032, USA.

† Correspondence to DV (dv2121@columbia.edu)

Abstract

In this brief communication we further discuss the relationship between the rate of protein evolution and the misfolding avoidance hypothesis (MAH), as it was originally proposed. We also discuss results of empirical tests of the MAH, and demonstrate the lack of any substantial correlation between protein stability, quantified by protein folding free energy, and protein evolutionary rates. Based on multiple available evidence, the MAH is unlikely to play any major role in explaining a strong negative correlation between protein abundance and the protein molecular clock.

We read with interest the recent manuscript "Protein melting temperature cannot fully assess whether protein folding free energy underlies the universal abundance-evolutionary rate correlation seen in proteins" by Rostam M. Razban recently published in Molecular Biology and Evolution [1]. This paper is related to our previous analyses, i.e. Plata *et al.* (2010) [2] and especially Plata and Vitkup (2017) [3]. Specifically, the manuscript discusses the absence of empirical support for the protein misfolding avoidance hypothesis (MAH) based on genome-wide melting temperature (Tm) measurements by Leuenberger *et al.* [4]. To avoid potential misunderstandings in the field, in this brief communication we address inaccuracies in the characterization of our previous work, and also comment more broadly on experimental tests of the MAH.

The original MAH proposed that the sequences of highly abundant proteins evolve slowly *primarily* due to increased selection against misfolded protein toxicity [5, 6]. The MAH can be tested by investigating: 1) whether cellular protein abundance positively correlates with protein stability [6, 7], and 2) whether the correlation between protein abundance and stability substantially affects the variability of protein evolutionary rates. As we previously demonstrated, careful re-analysis of the proteome-wide Tm measurements obtained by Leuenberger *et al.* [4] provides no support for the MAH in *E. coli* and three other investigated species [3].

The central message of Razban's analysis is that the absence of the expected relationship between Tm and protein abundance may be due to an imperfect correlation between Tm and protein stability (unfolding) free energies (ΔG), as well as experimental noise in Tm measurements [1]. While it is often possible to invoke sophisticated noise and error models to explain the absence of expected observations, this does not change the key scientific result that the available experimental data do not support a major role of the MAH in constraining evolution of highly expressed proteins. Notably, the MAH is also not supported by several direct experimental studies that assessed the fitness costs of protein expression and the costs of protein misfolding [2, 8].

Razban's manuscript states that our analysis provides weak support for the MAH in *E. coli*. This claim is not correct and, we believe, exemplifies a common and unfortunate confusion. As we specifically discussed [2, 3], the MAH cannot be validated simply by demonstrating weak correlations between protein abundance and stability. The MAH, at its core, is not about the stability of highly expressed proteins, but about a major effect of protein stability on the diversity of sequence constraints across cellular proteins. Thus, it is crucial to investigate whether protein stability, measured either using Tm or ΔG, accounts for any substantial fraction of the variance of evolutionary rates (molecular clock). Using partial correlation analyses between the protein molecular clock and abundance, while controlling for protein stability, we previously demonstrated that this is clearly not the case (see Table 1 in Plata and Vitkup [3]). We also showed, using genome-wide Tm measurements, the absence of a substantial direct correlation between protein stability and the protein molecular clock (see Figure 2 in Plata and Vitkup [3]). Similar results are also obtained for the correlation between molecular clock rates and ΔG measurements available in the ProTherm database (Figure 1). We note in this respect that proteins clearly need to be stable for performing their cellular functions, and protein stability does constrain sequence evolution [9]. Nevertheless, protein stability does not seem to have a major effect on the *variability* of evolutionary rates across cellular proteins.

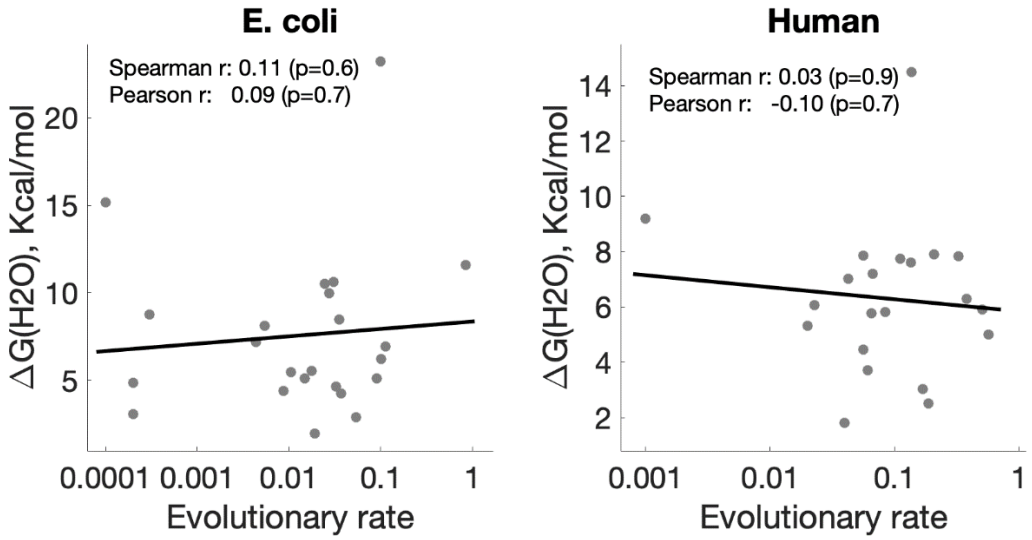

Figure 1. Correlations between protein evolutionary (molecular clock) rates and experimentally measured protein stability free energies (ΔG). **A.** *E. coli* (n=23) and **B.** *Homo sapiens* (n=21). The protein stability data were obtained from the ProTherm database. We considered the average stability (unfolding) free energies for wild type monomers measured at pH values between 4 and 9 and temperatures between 10 and 50 °C. We used evolutionary rates from Plata *et al.* [3].

Finally, although the feasibility of the MAH was demonstrated by computational simulations [6, 10, 11], biology is an empirical science, and the fidelity of proposed hypotheses should be ultimately determined by their agreement, or lack thereof, with available experimental data. Based on the preponderance of available evidence, the MAH is unlikely to play any major role in explaining a strong negative correlation between protein abundance and evolutionary rates. We note that major roles are also unlikely for several other effects, such as increasing protein solubility or the avoidance of non-functional interactions and binding, which was first demonstrated in *E. coli* [2] and then confirmed in yeast [12]. Most importantly, the search for the main factors contributing to the substantial variability of protein molecular clock rates must continue.

Acknowledgements

We would like to thank Purushottam Dixit for scientific discussions. This work was supported by the National Institute of General Medical Sciences grant R35GM131884 to Dennis Vitkup.

References

1. Razban, R.M. (2019). Protein melting temperature cannot fully assess whether protein folding free energy underlies the universal abundance-evolutionary rate correlation seen in proteins. Mol Biol Evol.
2. Plata, G., Gottesman, M.E., and Vitkup, D. (2010). The rate of the molecular clock and the cost of gratuitous protein synthesis. Genome Biol *11*, R98.

3. Plata, G., and Vitkup, D. (2017). Avoidance of toxic misfolding and protein stability do not explain the sequence constraints of highly expressed proteins. Mol Biol Evol.
4. Leuenberger, P., Ganscha, S., Kahraman, A., Cappelletti, V., Boersema, P.J., von Mering, C., Claassen, M., and Picotti, P. (2017). Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science *355*.
5. Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA *102*, 14338-14343.
6. Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341-352.
7. Zhang, J., and Yang, J.R. (2015). Determinants of the rate of protein sequence evolution. Nat Rev Genet *16*, 409-420.
8. Kafri, M., Metzl-Raz, E., Jona, G., and Barkai, N. (2016). The Cost of Protein Production. Cell Rep *14*, 22-31.
9. Dill, K.A., and Bromberg, S. (2011). Molecular driving forces : statistical thermodynamics in biology, chemistry, physics, and nanoscience, 2nd Edition, (London ; New York: Garland Science).
10. Serohijos, A.W., Rimas, Z., and Shakhnovich, E.I. (2012). Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep *2*, 249-256.
11. Yang, J.R., Zhuang, S.M., and Zhang, J. (2010). Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol *6*, 421.
12. Yang, J.R., Liao, B.Y., Zhuang, S.M., and Zhang, J. (2012). Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A *109*, E831-840.